MotionChain: Conversational Motion Controllers via Multimodal Prompts

Biao Jiang[1,2*], Xin Chen[2†], Chi Zhang[2], Fukun Yin[1], Zhuoyuan Li[1], Gang Yu[2], and Jiayuan Fan[1‡]

[1] Fudan University
[2] Tencent

This appendix provides several additional experiments (Sec. A), more qualitative results (Sec. B), model implementation details (Sec. C), evaluations of inference time (Sec. D), protocol for the motion conversation evaluation (Sec. E), details of motion representations (Sec. F), metric definitions (Sec. G).

A Additional Experiments

We conducted a comprehensive series of experiments to evaluate the efficacy of the proposed MotionChain models further. Specifically, we evaluate each specific comparison on text-to-motion (Sec. A.1), motion-to-text (Sec. A.2), and motion prediction (Sec. A.3) on the HumnaML3D [6] dataset. Additionally, we present an ablation study focusing on the effectiveness of our motion tokenizer (Sec. A.5) and the integration of motion tokens within the language model (Sec. A.6).

A.1 Comparisons on Text-to-Motion

The text-to-motion task showcases our MotionGPT model's capability in generating human-like movements based on textual inputs. Evaluations were performed on MotionChain against current state-of-the-art methods [6, 7, 10, 32, 34, 36], on the HumanML3D [6] dataset according to established metrics [6]. The evaluation results, featuring a 95% confidence interval from 20 runs, largely draw from data reported in the cited works. The comparative outcomes, summarized in Tab. 5, demonstrating MotionChain's competitive performance across numerous metrics.

A.2 Comparisons on Motion-to-Text

In the motion-to-text task, the goal is to generate descriptive text based on sequences of human motion. We evaluate the proposed MotionChain, contrasting it with TM2T [7] and MotionGPT [10] on the HumanML3D dataset and adhering to the evaluation metrics used in [7, 10]. Following [10], we leverages the original ground truth texts for evaluation, ensuring a more comprehensive assessment . Assessments in Tab. 6 demonstrate that MotionChain outperforms the recent methods in generating text descriptions of human motions on most benchmarks.

*Work done while Biao Jiang was a Research Intern with Tencent.
†Project lead.
‡Corresponding author.

Table 5: Comparison of text-to-motion on HumanML3D [6]. The empty MModality indicates *Real* motion is deterministic. *Pre-trained* and *Fine-tuned* indicate uniform motion-language pre-training and specific fine-tuning on this task. The arrows (→) indicate that closer to *Real* is desirable. **Bold** and underline indicate the best and the second best result on text-to-motion task.

Methods	RPrecision↑			FID↓	MMDist↓	Diversity→	MModality↑
	Top1	Top2	Top3				
Real	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	-
TM2T [7]	$0.424^{\pm.003}$	$0.618^{\pm.003}$	$0.729^{\pm.002}$	$1.501^{\pm.017}$	$3.467^{\pm.011}$	$8.589^{\pm.076}$	$\underline{2.424}^{\pm.093}$
T2M [6]	$0.457^{\pm.002}$	$0.639^{\pm.003}$	$0.740^{\pm.003}$	$1.067^{\pm.002}$	$3.340^{\pm.008}$	$9.188^{\pm.002}$	$2.090^{\pm.083}$
MotionDiffuse [37]	$0.491^{\pm.001}$	$\underline{0.681}^{\pm.001}$	$\underline{0.782}^{\pm.001}$	$0.630^{\pm.001}$	$3.113^{\pm.001}$	$9.410^{\pm.049}$	$1.553^{\pm.042}$
MDM [32]	$0.320^{\pm.005}$	$0.498^{\pm.004}$	$0.611^{\pm.007}$	$0.544^{\pm.044}$	$5.566^{\pm.027}$	$9.559^{\pm.086}$	$\mathbf{2.799}^{\pm.072}$
MLD [34]	$0.481^{\pm.003}$	$0.673^{\pm.003}$	$0.772^{\pm.002}$	$0.473^{\pm.013}$	$3.196^{\pm.010}$	$9.724^{\pm.082}$	$2.413^{\pm.079}$
T2M-GPT [36]	$0.491^{\pm.003}$	$0.680^{\pm.003}$	$0.775^{\pm.002}$	$\mathbf{0.116}^{\pm.004}$	$3.118^{\pm.011}$	$9.761^{\pm.081}$	$1.856^{\pm.011}$
MotionGPT [10]	$\underline{0.492}^{\pm.003}$	$\underline{0.681}^{\pm.003}$	$0.778^{\pm.002}$	$\underline{0.232}^{\pm.008}$	$\underline{3.096}^{\pm.008}$	$\mathbf{9.528}^{\pm.071}$	$2.008^{\pm.084}$
MotionChain	$\mathbf{0.504}^{\pm.003}$	$\mathbf{0.695}^{\pm.003}$	$\mathbf{0.790}^{\pm.003}$	$0.248^{\pm.009}$	$\mathbf{3.033}^{\pm.010}$	$\underline{9.470}^{\pm.075}$	$1.715^{\pm.066}$

Table 6: Comparison of motion captioning on HumanML3D [6]. The evaluation metrics follow [7], while we use the ground truth texts without pre-processing for linguistic metrics calculation. **Bold** indicate the best.

Methods	$\text{Length}_{\text{avg}}$↑	Bleu@1↑	Bleu@4↑	Rouge↑	Cider↑	BertScore↑
Real	12.75	-	-	-	-	-
TM2T [7]	10.67	**48.9**	7.00	38.1	16.8	32.2
MotionGPT [10]	**13.04**	48.2	12.47	37.4	29.2	32.4
MotionChain	12.37	48.1	**12.56**	**39.9**	**33.7**	**36.9**

A.3 Comparisons on Motion Completion.

In accordance with MotionGPT [10], we consider motion prediction as a collective task referred to as general motion completion. To assess the motion completion capability of MotionChain, we utilize a subset of the AMASS dataset [16], which consists solely of motion data. For the motion prediction task, we use only the initial 20% of the motion sequence as conditions. We evaluate MotionChain using the identical settings as outlined in [10]. The motion completion results of MotionChain, presented in Table 7, indicate that MotionChain achieves lower values in terms of ADE and FDE metrics. This implies that the mean and last-frame L2 distance between the ground truth and predicted motion are closer.

A.4 Evaluation on Multi-turn performance.

We evaluated performance as the number of motions and conversation turns increased, as shown in Fig. 5. (a) We split the GT motions into N equal-length sequences and had the models generate N times. (b) We assessed visual conditioned generation performance in different conversation rounds.

Table 7: Comparison of motion composition on HumanML3D. FID indicates motion quality and Diversity (DIV) for motion diversity within each condition. ADE and FDE are joints distance between generation and ground truth.

Methods	Motion Prediction				Motion In-between		
	FID ↓	Diversity↑	ADE↓	FDE↓	FID ↓	Diversity↑	ADE↓
Real	0.002	9.503	-	-	0.002	9.503	-
MDM [32]M	6.031	7.813	5.446	8.561	2.698	8.420	3.787
T2M-GPT [36]	2.056	8.635	6.161	8.302	-	-	-
MotionGPT [10]	**0.905**	8.972	4.745	6.040	**0.214**	**9.560**	3.762
MoMask [5]	2.546	9.044	3.514	**5.079**	0.548	9.691	**2.026**
MotionChain - small	1.607	8.172	5.162	6.859	0.634	9.099	3.514
MotionChain - base	1.053	8.802	4.388	5.401	0.325	8.821	2.939
MotionChain - large	1.004	**9.107**	**3.437**	5.213	0.239	8.853	2.624

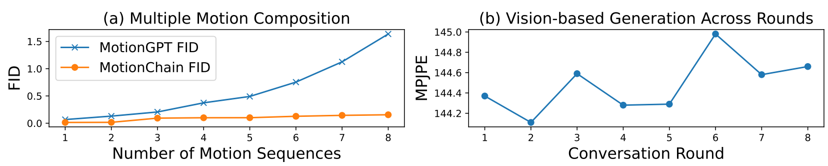

Fig. 5: Visualization of performance changes.

A.5 Ablation on Motion Tokenizer.

we conducted an ablation study on the motion tokenizer $\mathcal{V}$ of the MotionChain model, focusing specifically on the impact of varying the size K and dimension d of motion codebooks, and residual quantizer layers Q. Additionally, we benchmarked our VQ-VAE implementation against previous work [21, 24, 34], as shown in Tab. 8. This comparative analysis underscored the better performance of our VQ-VAE approach in terms of motion reconstruction accuracy. Through this comprehensive ablation study, in addition to the length limit of T5 series models, we thus identified parameters for the majority of our experiments as $Q = 4, K = 512, d = 1024$.

A.6 Ablation on Motion Tokens.

Subsequent to our analysis of motion codebooks, we shift focus to the strategy of sharing motion vocabularies V_m within the language model backbone. Specifically, we aim to explore the differences between sharing motion tokens across different quantization layers in the language model (LM) and not sharing them. For the LM codebooks, we design a baseline where motion tokens from different layers are shared, resulting in V_m newly added tokens. In another setting, where tokens are not shared, this results in $V_m \times Q$ newly added tokens. All other settings, such as the motion tokenizer, are kept the same. Our experiment shown in Tab. 9, grounded in the text-to-motion experiments conducted on the HumanML3D [6] dataset, reveals that the best performance is achieved when motion codes are not shared across the language model.

Table 8: Evaluation of our motion tokenizer on the motion part of HumanML3D [6] dataset. We follow MLD [34] to evaluate our VQ-VAE model $\mathcal{V}$: MPJPE and PAMPJPE are measured in millimeter. We evaluate FID and Diversity the same as Tab. 3. The baselines of VPoser-t [21] and ACTOR [24] are borrowed from MLD. K indicates the codebook size, d indicates the codebook dimension , Q indicates the Residual-VQ layers.

Method	Reconstruction			
	MPJPE↓	PAMPJPE↓	FID↓	DIV→
Real	-	-	0.002	9.503
VPoser-t [21]	75.6	48.6	1.430	8.336
ACTOR [24]	65.3	41.0	0.341	**9.569**
MLD-1 [34]	**54.4**	41.6	0.247	9.630
MotionGPT [10]	55.8	**40.1**	0.067	9.675
MotionChain	63.1	43.4	**0.014**	9.157
$Q=4, K=128, d=512$	71.8	51.2	0.037	9.098
$Q=4, K=256, d=512$	70.4	48.5	0.051	9.004
$Q=4, K=512, d=512$	69.5	46.5	**0.025**	9.015
$Q=4, K=1024, d=512$	**65.9**	**43.9**	0.041	**9.310**
$Q=2, K=512, d=512$	79.7	56.9	0.081	9.162
$Q=4, K=512, d=512$	69.5	46.5	0.025	9.015
$Q=8, K=512, d=512$	49.7	38.6	**0.025**	**9.213**
$Q=16, K=512, d=512$	**48.4**	**38.4**	0.026	9.075
$Q=4, K=512, d=128$	114.5	79.7	1.698	8.344
$Q=4, K=512, d=256$	83.9	59.7	0.560	8.782
$Q=4, K=512, d=512$	69.5	46.5	0.025	9.015
$Q=4, K=512, d=1024$	**63.1**	**43.4**	**0.014**	**9.157**

Table 9: Comparison of text-to-motion on HumanML3D [6]. The empty MModality indicates *Real* motion is deterministic. *Shared* indicate motion tokens from different layers are shared in the language model (LM), resulting in V_m newly added tokens. *Independent* indicates tokens are not shared which results in $V_m \times Q$ newly added tokens. The arrows (→) indicate that closer to *Real* is desirable. **Bold** and underline indicate the best and the second best result on text-to-motion task.

Methods	Motion Token Numbers	RPrecision↑			FID↓	MMDist↓	Diversity→	MModality↑
		Top1	Top2	Top3				
Real	-	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	-
Shared	V_m	$0.496^{\pm.003}$	$0.686^{\pm.003}$	$0.784^{\pm.002}$	$0.291^{\pm.012}$	$3.067^{\pm.011}$	$9.394^{\pm.075}$	$\mathbf{2.072}^{\pm.080}$
Independent	$V_m \times Q$	$\mathbf{0.504}^{\pm.003}$	$\mathbf{0.695}^{\pm.003}$	$\mathbf{0.790}^{\pm.003}$	$\mathbf{0.248}^{\pm.009}$	$\mathbf{3.033}^{\pm.010}$	$\mathbf{9.470}^{\pm.075}$	$1.715^{\pm.066}$

B Qualitative Results

We visualize our result gallery on motion conversations ($cf.$ Fig. 6) and some qualitative results on the comparison of text-to-motion ($cf.$ Fig. 7) and motion reasoning ($cf.$ Fig. 8).

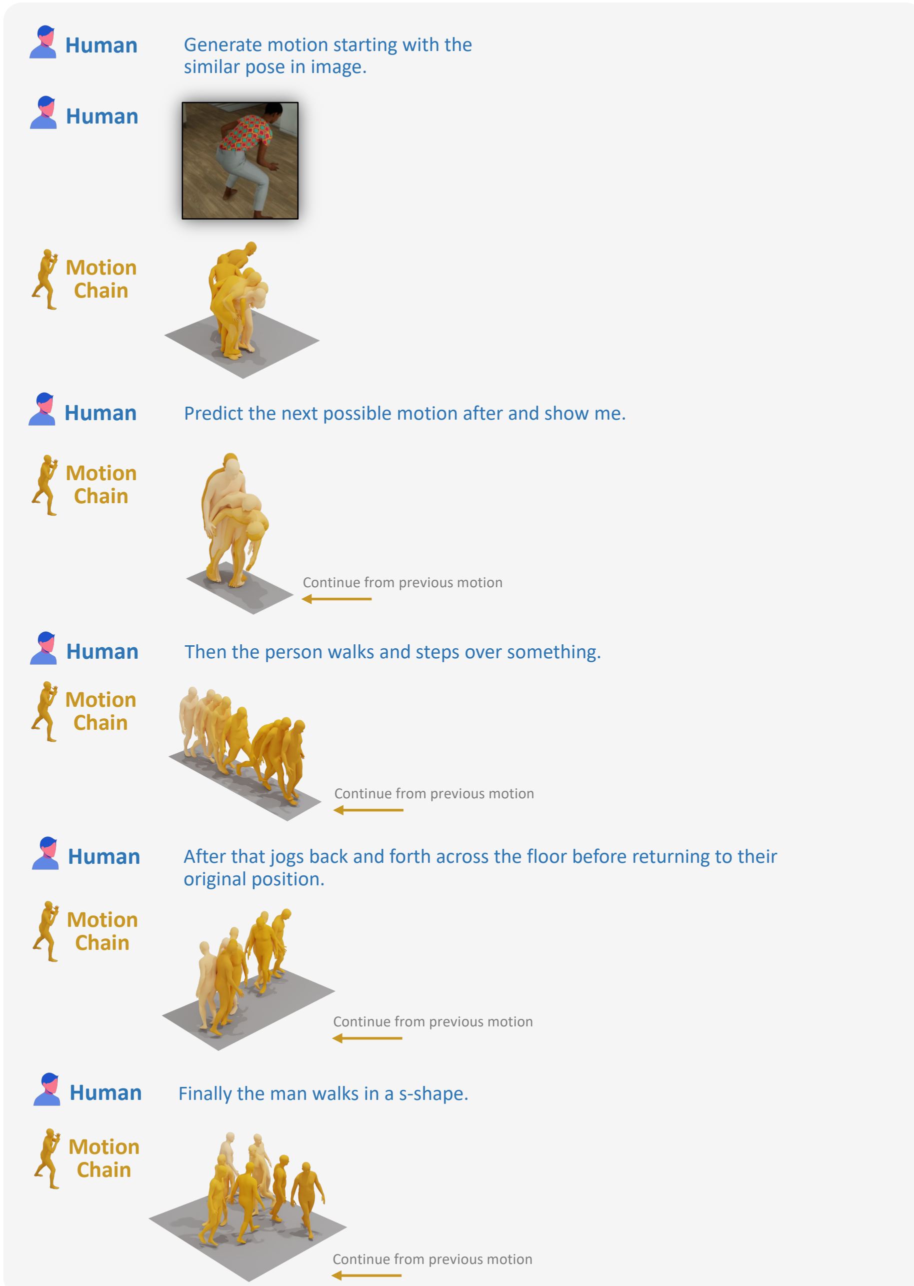

Fig. 6: The gallery showcases the results of our MotionChain model. The supervision of MotionChain is based on our conversational motion-language dataset (see Appendix E), which builds upon previous motion datasets [6, 27]. For a more dynamic visualization, we recommend referring to our supplemental video.

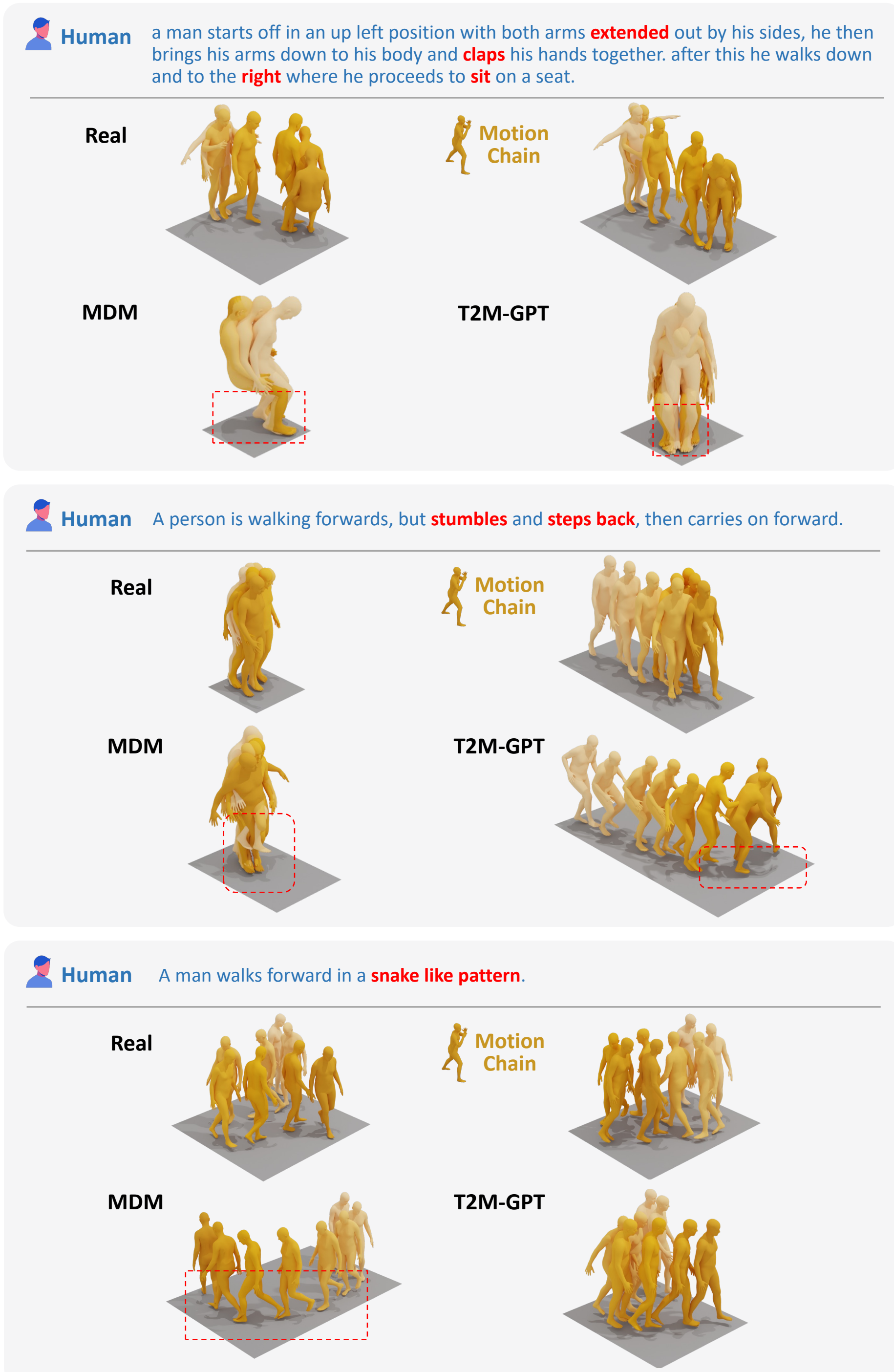

Fig. 7: Comparison of text-driven motion generation methods on the HumanML3D dataset [6]. In the visualizations, misaligned motions are highlighted with red words and boxes, while the characters are color-coded from light to dark to indicate the progression of time.

Human What is happening in the motion?

GT The person is swinging a golf club.

Motion Chain In the motion, a person is standing with their arms out from their sides, then they lower their arms and with their right hand swipe in front of them, left to right, up to down, then out to in.

Flan-T5-xl A person is swinging their arms as if they were hitting a golf ball.

Llama-7b Based on the given human motion, it appears that the person is practicing a golf swing. The bent posture and the holding of something with both hands suggest that the person is gripping a golf club, and the swinging of the arms is likely intended to mimic the motion of hitting a golf ball.

Vicuna-7b Based on the description provided, it appears that the person is performing a golf swing motion. In a golf swing, the person stands with their feet shoulder-width apart, bends slightly forward at the waist, and holds onto something with both hands. They then swing their arms back and forth, transferring their weight from their back foot to their front foot as they do so. The goal of the golf swing is to generate enough force and accuracy to hit the ball with the club.

Human What is the difference between a professional golfer's swing and an amateur golfer's swing?

GT Professional golfers typically have more consistent swings and are able to hit the ball with greater accuracy and power. They also tend to have more refined techniques and spend more time practicing their swing. Amateur golfers may have less consistent swings and may struggle with hitting the ball accurately or with enough force.

Motion Chain A professional golfer's swing will be more efficient and well-coordinated, with less wasted motion and more power and accuracy. An amateur golfer's swing will be more sloppy and uncontrolled, with more swinging arms and a more awkward gait.

Flan-T5-xl The difference between a professional golfer's swing and an amateur golfer's swing is the amount of force that is applied to the club.

Llama-7b The main difference between a professional golfer's swing and an amateur golfer's swing is the level of coordination, control, and power generated throughout the swing. Professional golfers typically have a more refined and consistent swing, with a smoother transition from backswing to downswing, better ball positioning, and more efficient use of their body weight and muscles. This results in faster swing speeds, more accurate ball striking, and longer drives. In contrast, amateur golfers may have a less controlled and less efficient swing, with a slower swing speed and more inconsistent ball striking.

Vicuna-7b The difference between a professional golfer's swing and an amateur golfer's swing is that the professional golfer's swing is more consistent, fluid, and powerful, while the amateur golfer's swing is often less controlled and lacks the same level of power and precision. Professional golfers have typically spent more time practicing and honing their swing, which allows them to execute it with greater accuracy and control. Additionally, professional golfers often have a more efficient and effective technique, which allows them to generate more power and accuracy with less effort.

Human How does the person's body position affect the swing?

GT The person's body position, including their stance and posture, can affect the direction and power of the swing. Proper body positioning can help the person hit the ball more accurately and with greater force.

Motion Chain The person's body position affects the swing by determining the direction and force of the swing. For example, if the person is standing with their feet shoulder-width apart, they will swing the golf club from left to right, while if they are standing with their feet shoulder-width apart, they will swing the golf club from right to left.

Flan-T5-xl The person's body position affects the swing by determining the angle of the swing.

Llama-7b The person's body position significantly affects the swing in the following ways:

1. Posture: A stable and balanced posture is essential for generating power and control in the swing. A professional golfer will typically maintain a neutral spine angle, with a slight forward tilt at address, while an amateur may lean back or forth excessively.
2. Clubhead speed: The position of the body can influence the speed at which the clubhead travels. A more compact swing with a shorter backswing and longer follow-through can generate faster clubhead speed, while an over-the-top swing may result in slower speeds.
3. Ball position: The position of the body can affect the positioning of the ball in the stance. A professional golfer will typically position the ball slightly forward of the center of their stance, while an amateur may put the ball too far back or too far forward.
4. Clubface alignment: The position of the body can influence the alignment of the clubface at address. A professional golfer will typically align the clubface parallel to the target line, while an amateur may position the clubface closed or open.

In summary, the person's body position in the swing has a significant impact on the swing's power, control, and accuracy.

Vicuna-7b The person's body position affects the swing of the golf club in several ways. Firstly, the person's stance, which is the position of their feet and legs, can affect the power and accuracy of the swing. A wider stance can provide more stability and power, while a narrower stance can increase control and accuracy.

Secondly, the person's grip on the golf club can also affect the swing. A loose grip can allow for more control and accuracy, while a tight grip can increase power but reduce control.

Thirdly, the person's posture and alignment can also affect the swing. Good posture and alignment can help the person maintain control and accuracy throughout the swing, while poor posture and alignment can lead to inconsistencies and errors.

Overall, the person's body position plays a crucial role in the swing of the golf club, and proper technique and practice can help the person achieve the desired body position for a successful swing.

Human Is there any equipment or tools used in this motion?

GT No, this motion does not require any equipment or tools.

Motion Chain No, there are no equipment or tools used in this motion. It is a simple body movement.

Flan-T5-xl No, there is no equipment or tools used in this motion.

Llama-7b Yes, there is one piece of equipment or tool used in the motion of \"person slides forward diagonally with right leg\":\n\n* Skates or ice skates

Vicuna-7b No, there is no equipment or tools used in this motion. The person is simply sliding forward diagonally with their right leg.

Human What might happen after this motion?

GT It depends on the context, but the person could continue moving forward or change direction.

Motion Chain Again, it's hard to say without more context. The person could continue moving in the same direction, or they might change direction again.

Flan-T5-xl You might be able to get up.

Llama-7b Based on the human motion provided, it is likely that the person will continue to move forward after the diagonal slide with their right leg. They may:

A) Take a step forward with their left foot to maintain their forward momentum.

B) Change direction and turn to the right or left.

C) Come to a stop or slow down.

D) Perform another motion, such as bending or stretching.

Vicuna-7b It is not possible for me to accurately answer your question without more context. The motion you describe is a simple action, but it could have many different possible outcomes depending on the situation and the person performing the motion. For example, the person might continue to slide forward diagonally with their right leg, or they might stop or change direction. Without more information, it is not possible to accurately predict what might happen after this motion.

Human Is there any specific type of footwear that would be best for performing this motion.

GT It would be beneficial for the person to wear shoes with good traction and support to prevent slipping or injury.

Motion Chain It's possible that this motion could be appropriate for someone who needs to perform this motion that requires a lot of balance, such as law enforcement or military. However, it would depend on the specific type of dance or athletic shoes.

Flan-T5-xl You might be able to get up.

Llama-7b Yes, the type of footwear that would be best for performing the motion of taking a step to the right would depend on the specific context and the individual's footwear needs. However, in general, a person may find it more comfortable and supportive to perform this motion in footwear with the following characteristics:

1. Good arch support: A step to the right requires the foot to absorb the impact of the step and transfer the weight to the right foot. Footwear with good arch support can help reduce the risk of foot fatigue and injury.

2. Wide toe box: A wide toe box can provide adequate room for the toes to move and flex, allowing the foot to move naturally during the step.

3. Traction: The footwear should have good traction on the ground to ensure stability and balance during the step. In summary, the best footwear for performing the motion of taking a step to the right would depend on the individual's foot shape, the surface they are stepping on, and the purpose of the step.

Vicuna-7b There is no specific type of footwear that is best for performing the motion of "a person making a step to the right." The type of footwear that is best for this motion would depend on various factors such as the surface on which the person is stepping, the weight and size of the person, and the purpose of the motion. For example, if the person is stepping on a rough or uneven surface, they may want to wear shoes with good traction and support. If the person is stepping for exercise or sports, they may want to wear shoes that provide good cushioning and support. Ultimately, the best type of footwear for performing the motion of "a person making a step to the right" would depend on the specific circumstances and needs of the person.

Fig. 8: Comparison on motion reasoning question-answer. The MotionChain is trained on our conversation dataset based on HumanML3D [6]. The results demonstrate that our MotionChain shows promising text and motion understanding.

C Implementation Details

We provide detailed explanations regarding the implementation details of motion composition (Sec. C.1), and the image tokenizer (Sec. C.2).

C.1 Details of Temporal Motion Compoistion

To investigate the temporal motion composition abilities of the MotionChain model, we conduct a pair actions composition experiment on the BABEL dataset [27], following the methodology of TEACH [2]. For simplicity, we consider pairs of actions, but it is important to note that MotionChain can handle sequences of actions/motions of any length. During training, in cases where there is segment overlap, we evenly distribute the overlapping frames between the two segments that form the pair. It is worth mentioning that the majority of the pair data (approximately 70 %) is generated through overlapping segments rather than transitions. In the event of a transition, we concatenate the transition with the second segment. Instead of training a MotionChain model from scratch on the BABEL dataset [27], we utilize a pre-trained MotionChain model obtained from HumanML3D [6]. Subsequently, we convert the motion data in the BABEL dataset [27] into the format used in HumanML3D [6], and then fine-tune the MotionChain model on the BABEL dataset [27] using prompts that incorporate memory, as demonstrated below:

> $X_{\text{system-message}}$
> USER: Please assume the role of an Human Motion Language translator. I will use English, you should translate it, and respond in Human Motion Language. My first request is "<label1>"
> ASSISTANT: <motion1>
> USER: Please assume the role of a Human Motion Language translator. I will use English, you should translate it, and respond in Human Motion Language. In the last round I asked you to translate "<label1>", and your answer is <motion1>. Now my second request is "<label2>"
> ASSISTANT: <motion2>

For comparison with TEACH [2], we employed the TEACH model that was pre-trained on the BABEL dataset [27] to generate motion samples 20 times on the validation set. Subsequently, we converted the generated motion, originally in SMPL-H format [14], into the HumanML3D format.

We also examine the influence of various motion composition mechanisms on the generated complete motion sequences, as presented in Table 3. The "Independent" mechanism refers to the direct concatenation of independently generated motion sequences without any additional processing. On the other hand, the "Tokens-joint" mechanism involves concatenating motion tokens and decoding them using the VQ decoder, which results in a more coherent and natural sequence of movements.

C.2 Details of Image Tokenzier

We explore three different architectural designs for image tokenizers:

(a) *MLP*: In this design, we connect the frozen vision encoder CLIP ViT-L/14 [28] to the language model using a linear layer. The output of the vision encoder is projected to the same dimension as the word embeddings of the language model and is inserted before the text or motion token embeddings.

(b) *Perceiver*: This design incorporates a perceiver module with a similar architecture to Flamingo [1]. The perceiver module includes a transformer that receives a predefined number of latent input queries. These queries are then projected to the same dimension as the word embeddings of the language model and are inserted before the text or motion token embeddings. Details of architecture is presented in Tab. 10.

(c) *Q-former*: In this design, we directly utilize the pre-trained Q-former from BLIP-2 [12] to align visual inputs with the language model. The Q-former is frozen throughout the entire training process.

Table 10: Architecture of our vision perceiver

```
(0): PerceiverResampler(
  (layers): ModuleList(
    (0-5): 6 x ModuleList(
      (0): PerceiverAttention(
        (norm_media): LayerNorm((1024,), eps=1e-05, elementwise_affine=True)
        (norm_latents): LayerNorm((1024,), eps=1e-05, elementwise_affine=True)
        (to_q): Linear(in_features=1024, out_features=512, bias=False)
        (to_kv): Linear(in_features=1024, out_features=1024, bias=False)
        (to_out): Linear(in_features=512, out_features=1024, bias=False) )
      (1): Sequential(
        (0): LayerNorm((1024,), eps=1e-05, elementwise_affine=True)
        (1): Linear(in_features=1024, out_features=4096, bias=False)
        (2): GELU(approximate='none')
        (3): Linear(in_features=4096, out_features=1024, bias=False) ) ) )
  (norm): LayerNorm((1024,), eps=1e-05, elementwise_affine=True) )
(1): Linear(in_features=1024, out_features=768, bias=True)
```

C.3 Details of Motion Similarities

As mentioned in Sec. 3.1, we employ TMR [25] for categorizing motions from the dataset into varying similarity levels. Here we define motion similarities sij > 0.8 are high and 0.6 < sij < 0.8 are medium

C.4 Details of Model Architecture

We provide model details in Fig. 9.

D Inference time

We conducted a study to evaluate the inference time of our MotionChain model, which utilizes an auto-regressive approach for motion generation. To assess the time costs, we measured the Frames Per Second (FPS) on a single Tesla V100 GPU with a batch size of one. It is important to note that the frame generation rate of MotionChain, even without specific engineering optimizations, surpasses the ground-truth frame rate in text-motion pair datasets [6, 13, 27], highlighting its capability to support real-time motion animation applications.

E Evaluation Protocols on the Motion Conversation.

We propose a protocol to evaluate our Multi-turn Multi-modal model, MotionChain, on various motion-language generation tasks. While MotionGPT [10] utilized previous text-motion pair datasets [6, 16, 26] to create an instruction motion-language dataset comprising 14 core tasks with numerous instruction templates, these tasks lack analysis

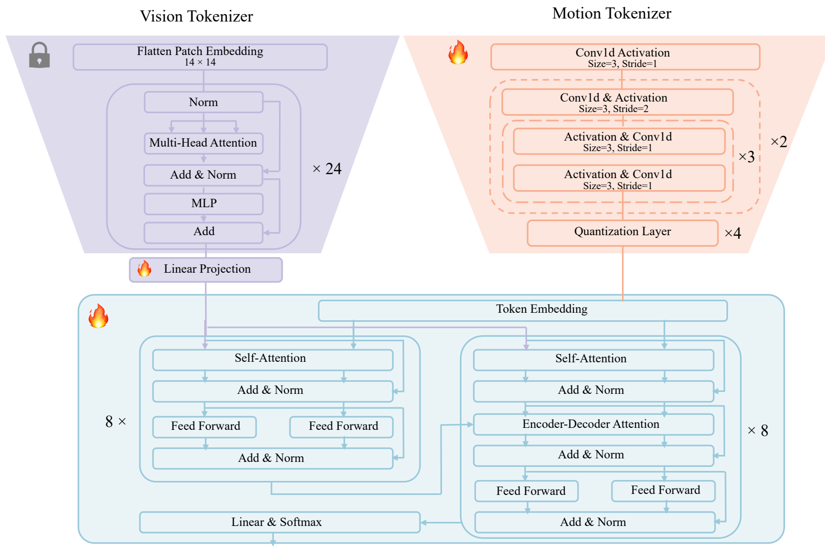

Fig. 9: MotionChain base architecture.

Table 11: The inference time costs of text-driven motion generation by evaluating the Frames Per Second (FPS), which is obtained by averaging the number of frames generated per second. We present the time costs for various model sizes and observe that, under the same 1 Tesla V100, smaller model sizes achieve faster FPS.

Models	Backbone	Parameters	FPS
MotionChain-small	Flan-T5-small	401 M	136.7
MotionChain-base	Flan-T5-Base	573 M	74.99
MotionChain-large	Flan-T5-Large	1.1 B	39.18

of human motion and are limited to single-turn generation without contextual memory. To overcome this limitation, we introduce motion reasoning and motion editing tasks that leverage contextual information. Initially, we manually provide ChatGPT [18, 19] with a few examples along with corresponding textual descriptions of the motions in the datasets, and then we let it generate the motion analysis (refer to Fig. 10). Additionally, using a pre-trained text-motion retrieval model, TMR [25], we retrieve motions from the dataset with high and middle similarities. We collect captions for motion pairs with middle similarity and employ ChatGPT [18, 19] to generate motion editing instructions that can transform one motion into another. Furthermore, we manually construct highly similar motion pairs for motion length editing tasks based on their respective lengths. By randomly combining these single-turn generation tasks, we can create a dialog format. The resulting tasks, along with diverse prompt instructions, are presented in Tab. 12. We provide token and round length statistics of proposed multi-turn datasets in Fig. 11. We will release the pre-processed dataset.

Prompt

```
You are an AI visual assistant, and you are seeing a motion.
What you see are provided with some sentences, describing the same motion you are looking at.
The motion content you are seeing is provided as following:

<Motion_Caption>

Design a conversation between you and a person asking about this motion.
In conversations you should indicate who saids using "User:","AI:" in the beggining but these two words
do not occur in sentences.
The answers should be in a tone that a AI visual assistant is seeing the motion and answering the
question.
Ask diverse questions and give corresponding answers, and questions are asking about the content of the
motion.
Questions the person will ask is related but not limited to following:
1. Asking the possible scenario of the motion happening.
2. Asking about what will happen after the motion or what have happened before the motion.
3. Asking what job the motion subject most likely to do.
4. Asking what tools or equipments are used during the motion, etc.
You also have to take some rigorous logic inference about the motion content and design other reasoning
questions-answers not limited to above examples.
Questions should not be yes-no questions but wh-questions.
The questions-answers you should design at least 9.
Provide detailed answers when answer questions, for example, give detailed examples or reasoning steps
to make the content more convincing and well-organized.
It is better to keep the questions concise, but the answers you design is allowed to be longer.
```

Fig. 10: The dedicated ChatGPT prompt for facilitating the collection of motion question-answer pairs. Our primary goal was to encompass a wide range of topics, including motion physics and motion analysis. By utilizing this prompt, our aim was to enable ChatGPT to generate high-quality questions, thereby making a valuable contribution to the development of a comprehensive motion question-answer dataset.

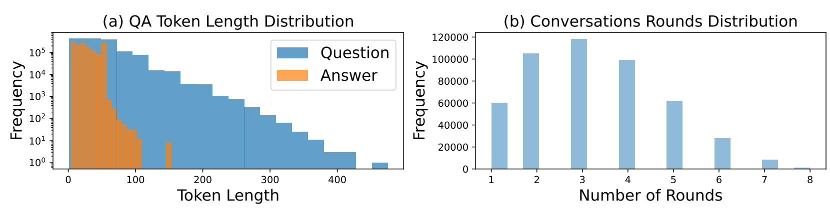

Fig. 11: Statistics of token length and rounds of our dataset.

F Motion Representations

We summarize two kinds of motion representations as follows.

HumanML3D Format [6] introduces a motion representation $x^{1:L}$ that draws inspiration from motion features in character control [23, 30, 31]. This representation, which contains redundant information, is well-suited for neural models, particularly variational autoencoders. Specifically, the i-th pose x^i is defined by a tuple consisting of the root angular velocity $\dot{r}^a \in \mathbb{R}$ along the Y-axis, root linear velocities $(\dot{r}^x, \dot{r}^z \in \mathbb{R})$ on the XZ-plane, root height $r^y \in \mathbb{R}$, local joint positions $\mathbf{j}^p \in \mathbb{R}^{3N_j}$, velocities $\mathbf{j}^v \in \mathbb{R}^{3N_j}$, and rotations $\mathbf{j}^r \in \mathbb{R}^{6N_j}$ in root space. Additionally, it includes binary foot-ground contact features $\mathbf{c}^f \in \mathbb{R}^4$ obtained by thresholding the heel and toe joint velocities. Here, N_j represents the number of joints, yielding the following representation:

$$x^i = \{\dot{r}^a, \dot{r}^x, \dot{r}^z, r^y, \mathbf{j}^p, \mathbf{j}^v, \mathbf{j}^r, \mathbf{c}^f\}. \tag{6}$$

SMPL-based Format [14] is a widely used parametric human model, SMPL [14], and its variants [22, 29], which propose motion parameters θ and shape parameters β.

Table 12: A few examples of prompt templates used in our standardized motion conversation evaluation protocol.

Task	Input	Output
Text-to-Motion	Show me a sequence of movements that illustrates [caption]. Demonstrate a motion that symbolizes the input: [caption]. I need a human motion that represents [caption].	[motion]
Text-to-Motion w/ length	Please generate a motion that is around [frames] frames long for the caption: [caption]. Generate a motion that lasts for [seconds] seconds, and captures the essence of [caption].	[motion]
Motion-Length-Editting	Extend the duration of the motion provided. Reduce the duration of the motion without losing its main characteristics and precision.	[motion]
Length-to-Motion	I want to see a motion that lasts for [frames] frames. Show me a motion that has a duration of [seconds] seconds.	[motion]
Radnom Motion	Just show me a moving human. Produce motions that are not planned or choreographed..	[motion]
Motion-to-Text	Provide a description of the motion shown in [motion] using natural language. Provide a text-based explanation of what is happening in [motion].	[caption]
Motion-to-Text w/ length	Generate a text summary for the [motion] that takes [frames] seconds to complete. Describe the movement exhibited in [motion] that is shown for a length of [seconds] seconds?	[caption]
Motion-to-Length	How long does [motion]'s poses last in seconds?? Calculate the second duration for [motion]'s body movements in seconds?	There are [frames] frames in the motion. The motion lasts for [seconds] seconds.
Caption-to-Length	Predict the anticipated frame duration for the motion that corresponds to [caption]? Guess the second count required for the motion represented by [caption].	The duration is estimated to be around [frames] frames. The motion has a length of [seconds] seconds.
Length-to-Caption	Given the [frames] frames of the motion, what are some possible actions that could be taken? [seconds] is the number of motion seconds, generate the motion description:	[caption]
Random Caption	Depict a motion as like you have seen it. Describe the motion of someone randomly.	[caption]
Motion-Reasoning	Can you tell me what muscles are being used during this motion?	This motion primarily targets the quadriceps, hamstrings, glutes, and core muscles. It also engages the shoulders and upper back muscles while raising the arms.
	What could be the reason for the person not swinging their arms while walking?	There could be various reasons for this, such as the person carrying something heavy or trying to maintain a certain posture while walking.

The rotation vectors $\theta \in \mathbb{R}^{3\times 23+3}$ represent the rotations of joints and the root, while β represents the weights for linear blended shapes. This representation is commonly employed in markerless motion capture [3, 9, 11]. By including the global translation r, the representation is formulated as:

$$x^i = \{r, \theta, \beta\}. \tag{7}$$

G Metric Definitions

In the following section, we present additional details regarding the evaluation metrics.

Linguistic Quality. To evaluate motion question-answer tasks, we employ linguistic metrics that assess the degree of alignment between the generated results and the ground-truth labels. These metrics include BLUE [20], Rouge citelin2004rouge, Cider [33], and BertScore [38]. For detailed information, please refer to the respective papers associated with each metric.

Motion Quality. The Frechet Inception Distance (FID) serves for evaluating the distribution similarity between generated and real motions. It is calculated using a suitable feature extractor [6, 8, 24] specific to each dataset. Additionally, we employ popular metrics in motion capture [3, 11, 17], such as MPJPE and PAMPJPE [4], to measure global and local errors in millimeters. To assess temporal quality, we utilize the Acceleration Error (ACCL). Furthermore, in line with previous motion prediction studies [15, 35, 39], we define the Average Displacement Error (ADE) as the average L2 distance between the ground truth and predicted motion for the entire sequence. The

Final Displacement Error (FDE) is calculated as the L2 distance between the ground truth and predicted motion in the last frame.

Motion Diversity. Following previous studies [7, 8, 32], we employ two metrics, Diversity (DIV) and MultiModality (MM), to evaluate the variability of motion across the entire dataset and the diversity of generated motion within each text input, respectively. To assess Diversity, the generated motions are randomly divided into two equal-sized subsets, and the Diversity metric is computed as the average distance between the motions in these subsets. For MultiModality evaluation, a set of text descriptions is randomly sampled from the available descriptions. Each text description is then replicated m times for motion generation, and the MultiModality metric is defined as the average distance between the motions generated from the same text description.

Condition Matching. HumanML3D [6] and TMR [25] provide motion/text feature extractors that generate geometrically coherent features for aligned text-motion pairs and vice versa. Within this feature space, we assess the motion-retrieval precision (R Precision) by combining the generated motion with 31 mismatched motions and calculating the top-1/2/3 matching accuracy between the text and motion. Additionally, we measure the Multi-modal Distance (MM Dist), which quantifies the distance between the generated motions and the corresponding text..

Time Costs. To assess the computational efficiency of our models, particularly the inference efficiency, we measure the average Frames Per Second (FPS) during motion generation. Specifically, we calculate the FPS on the test set of HumanML3D [6], with a batch size of one, while excluding the time required for model and dataset loading.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems **35**, 23716–23736 (2022)
2. Athanasiou, N., Petrovich, M., Black, M.J., Varol, G.: Teach: Temporal action compositions for 3d humans. In: International Conference on 3D Vision (3DV) (September 2022)
3. Chen, X., Pang, A., Yang, W., Ma, Y., Xu, L., Yu, J.: Sportscap: Monocular 3d human motion capture and fine-grained understanding in challenging sports videos. International Journal of Computer Vision **129**(10), 2846–2864 (2021)
4. Gower, J.C.: Generalized procrustes analysis. Psychometrika **40**(1), 33–51 (1975)
5. Guo, C., Mu, Y., Javed, M.G., Wang, S., Cheng, L.: Momask: Generative masked modeling of 3d human motions. arXiv preprint arXiv:2312.00063 (2023)
6. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5152–5161 (June 2022)
7. Guo, C., Zuo, X., Wang, S., Cheng, L.: Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In: ECCV (2022)
8. Guo, C., Zuo, X., Wang, S., Zou, S., Sun, Q., Deng, A., Gong, M., Cheng, L.: Action2motion: Conditioned generation of 3d human motions. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2021–2029 (2020)
9. He, Y., Pang, A., Chen, X., Liang, H., Wu, M., Ma, Y., Xu, L.: Challencap: Monocular 3d capture of challenging human performances using multi-modal references. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11400–11411 (2021)
10. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: Motiongpt: Human motion as a foreign language. arXiv preprint arXiv:2306.14795 (2023)
11. Kocabas, M., Athanasiou, N., Black, M.J.: Vibe: Video inference for human body pose and shape estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
12. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
13. Lin, J., Zeng, A., Lu, S., Cai, Y., Zhang, R., Wang, H., Zhang, L.: Motion-x: A large-scale 3d expressive whole-body human motion dataset. arXiv preprint arXiv:2307.00818 (2023)
14. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM Trans. Graph. **34**(6), 248:1–248:16 (Oct 2015)
15. Ma, H., Li, J., Hosseini, R., Tomizuka, M., Choi, C.: Multi-objective diverse human motion prediction with knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8161–8171 (2022)
16. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
17. von Marcard, T., Henschel, R., Black, M., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: European Conference on Computer Vision (ECCV) (sep 2018)
18. OpenAI: Gpt-4 technical report (2023)
19. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems **35**, 27730–27744 (2022)

20. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
21. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2019)
22. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 10975–10985 (Jun 2019), `http://smpl-x.is.tue.mpg.de`
23. Peng, X.B., Ma, Z., Abbeel, P., Levine, S., Kanazawa, A.: Amp: Adversarial motion priors for stylized physics-based character control. ACM Trans. Graph. **40**(4) (Jul 2021). `https://doi.org/10.1145/3450626.3459670`, `http://doi.acm.org/10.1145/3450626.3459670`
24. Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3D human motion synthesis with transformer VAE. In: International Conference on Computer Vision (ICCV) (2021)
25. Petrovich, M., Black, M.J., Varol, G.: Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. arXiv preprint arXiv:2305.00976 (2023)
26. Plappert, M., Mandery, C., Asfour, T.: The kit motion-language dataset. Big Data **4**(4), 236–252 (dec 2016). `https://doi.org/10.1089/big.2016.0028`, `http://dx.doi.org/10.1089/big.2016.0028`
27. Punnakkal, A.R., Chandrasekaran, A., Athanasiou, N., Quiros-Ramirez, A., Black, M.J.: BABEL: Bodies, action and behavior with english labels. In: Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 722–731 (Jun 2021)
28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
29. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) **36**(6) (Nov 2017)
30. Starke, S., Mason, I., Komura, T.: Deepphase: periodic autoencoders for learning motion phase manifolds. ACM Transactions on Graphics (TOG) **41**(4), 1–13 (2022)
31. Starke, S., Zhang, H., Komura, T., Saito, J.: Neural state machine for character-scene interactions. ACM Trans. Graph. **38**(6), 209–1 (2019)
32. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Bermano, A.H., Cohen-Or, D.: Human motion diffusion model. arXiv preprint arXiv:2209.14916 (2022)
33. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015)
34. Xin, C., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, J., Yu, G.: Executing your commands via motion diffusion in latent space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2023)
35. Yuan, Y., Kitani, K.: Dlow: Diversifying latent flows for diverse human motion prediction. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. pp. 346–364. Springer (2020)
36. Zhang, J., Zhang, Y., Cun, X., Huang, S., Zhang, Y., Zhao, H., Lu, H., Shen, X.: T2m-gpt: Generating human motion from textual descriptions with discrete representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
37. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. arXiv preprint arXiv:2208.15001 (2022)

38. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019)
39. Zhang, Y., Black, M.J., Tang, S.: We are more than our joints: Predicting how 3d bodies move. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3372–3382 (2021)